Active Vision for Sociable Robots

Cynthia Breazeal, Aaron Edsinger, Paul Fitzpatrick, and Brian Scassellati

Abstract—
In 1991, Ballard [1] described the implications of having a visual system that could actively position the camera coordinates in response to physical stimuli. In humanoid robotic systems, or in any animate vision system that interacts with people, social dynamics provide additional levels of constraint and provide additional opportunities for processing economy. In this paper, we describe an integrated visual-motor system that has been implemented on a humanoid robot to negotiate the robot's physical constraints, the perceptual needs of the robot's behavioral and motivational systems, and the social implications of motor acts.

Keywords— **Active vision, robots, social interaction with humans.**

I. Introduction

Animate vision introduces requirements for real-time processing, removes simplifying assumptions of static camera systems, and presents opportunities for simplifying computation. This simplification arises through situating perception in a behavioral context, by providing for opportunities to learn flexible behaviors, and by allowing the exploitation of dynamic regularities of the environment [1]. These benefits have been of critical interest to a variety of humanoid robotics projects [2], [3], [4], [5], and to the robotics and AI communities as a whole. On a practical level, the vast majority of these systems are still limited by the complexities of perception and thus focus on a single aspect of animate vision or concentrate on the integration of two well-known systems. On a theoretical level, existing systems often do not benefit from the advantages that Ballard proposed because of their limited scope.

In humanoid robotics, these problems are particularly evident. Animate vision systems that provide only a limited set of behaviors (such as supporting only smooth pursuit tracking) or that provide behaviors on extremely limited perceptual inputs (such as systems that track only very bright light sources) fail to provide a natural interaction between human and robot. We propose that in order to allow realistic human-machine interactions, an animate vision system must address a set of *social constraints* in addition to the other issues that classical active vision has addressed.

It is useful to view social constraints not as limitations, but opportunities. They induce a natural "vocabulary" to make the robot's behavior and state readable by a human. This in turn can provide a framework for the robot to negotiate a change in a human's behavior. For example, section XI-A discusses a procedure the robot can use to control the "inter-personal" distance a human assigns to it. Having this control means that in social situations the robot can get quite far with a simple vision system that is tuned to a particular distance.

Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA. E-mail: cynthia,edsinger,paulfitz,scaz@ai.mit.edu

II. Social constraints

For robots and humans to interact meaningfully, it is important that they understand each other enough to be able to shape each other's behavior. This has several implications. One of the most basic is that robot and human should have at least some overlapping perceptual abilities. Otherwise, they can have little idea of what the other is sensing and responding to. Vision is one important sensory modality for human interaction, and the one we focus on in this article. We endow our robots with visual perception that is human-like in its physical implementation.

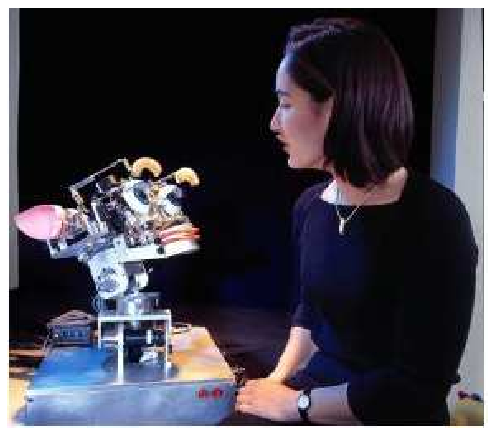

Fig. 1. Kismet, a robot capable of conveying intentionality through facial expressions and behavior. Here, the robot's physical state expresses attention to and interest in the human beside it. Another person – for example, the photographer – would expect to have to attract the robot's attention before being able to influence its behavior.

Similarity of perception requires more than similarity of sensors. Not all sensed stimuli are equally behaviorally relevant. It is important that both human and robot find the same types of stimuli salient in similar conditions. Our robots have a set of perceptual biases based on the human pre-attentive visual system. These biases can be modulated by the motivational state of the robot, making later perceptual stages more behaviorally relevant. This approximates the top-down influence of motivation on the bottom-up pre-attentive process found in human vision.

Visual perception requires high bandwidth and is computationally demanding. In the early stages of human vision, the entire visual field is processed in parallel. Later computational steps are applied much more selectively, so that behaviorally relevant parts of the visual field can be processed in greater detail. This mechanism of visual attention is just as important for robots as it is for humans, from the same considerations of resource allocation. The existence of visual attention is also key to satisfying the expectations of humans concerning what can and cannot be perceived visually. We have implemented a context-dependent attention system that goes some way towards this.

Human eye movements have a high communicative value. For example, gaze direction is a good indicator of the locus of visual attention. Knowing a person's locus of attention reveals what that person currently considers behaviorally relevant, which is in turn a powerful clue to their intent. The dynamic aspects of eye movement, such as staring versus glancing, also convey information. Eye movements are particularly potent during social interactions, such as conversational turn-taking, where making and breaking eye contact plays an important role in regulating the exchange. We model the eye movements of our robots after humans, so that they may have similar communicative value.

Our hope is that by following the example of the human visual system, the robot's behavior will be easily understood because it is analogous to the behavior of a human in similar circumstances (see Figure 1). For example, when an anthropomorphic robot moves its eyes and neck to orient toward an object, an observer can effortlessly conclude that the robot has become interested in that object. These traits lead not only to behavior that is easy to understand but also allows the robot's behavior to fit into the social norms that the person expects.

There are other advantages to modeling our implementation after the human visual system. There is a wealth of data and proposed models for how the human visual system is organized. This data provides not only a modular decomposition but also mechanisms for evaluating the performance of the complete system. Another advantage is robustness. A system that integrates action, perception, attention, and other cognitive capabilities can be more flexible and reliable than a system that focuses on only one of these aspects. Adding additional perceptual capabilities and additional constraints between behavioral and perceptual modules can increase the relevance of behaviors while limiting the computational requirements [6]. For example, in isolation, two difficult problems for a visual tracking system are knowing what to track and knowing when to switch to a new target. These problems can be simplified by combining the tracker with a visual attention system that can identify objects that are behaviorally relevant and worth tracking. In addition, the tracking system benefits the attention system by maintaining the object of interest in the center of the visual field. This simplifies the computation necessary to implement behavioral habituation. These two modules work in concert to compensate for the deficiencies of the other and to limit the required computation in each.

III. Physical form

Currently, the most sophisticated of our robots in terms of visual-motor behavior is Kismet. This robot is an active vision head augmented with expressive facial features (see Figure 2). Kismet is designed to receive and send human-like social cues to a caregiver, who can regulate its environment and shape its experiences as a parent would for a child. Kismet has three degrees of freedom to control gaze direction, three degrees of freedom to control its neck, and fifteen degrees of freedom in other expressive components of the face (such as ears and eyelids). To perceive its caregiver Kismet uses a microphone, worn by the caregiver, and four color CCD cameras. The positions of the neck and eyes are important both for expressive postures and for directing the cameras towards behaviorally relevant stimuli.

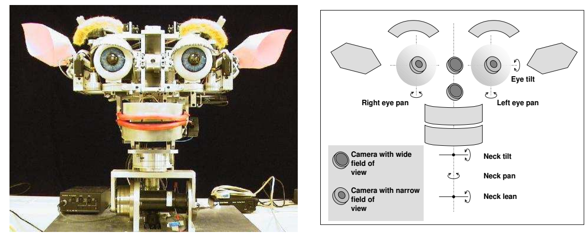

Fig. 2. Kismet has a large set of expressive features – eyelids, eyebrows, ears, jaw, lips, neck and eye orientation. The schematic on the right shows the degrees of freedom relevant to visual perception (omitting the eyelids!). The eyes can turn independently along the horizontal (pan), but turn together along the vertical (tilt). The neck can turn the whole head horizontally and vertically, and can also crane forward. Two cameras with narrow "foveal" fields of view rotate with the eyes. Two central cameras with wide fields of view rotate with the neck. These cameras are unaffected by the orientation of the eyes.

The cameras in Kismet's eyes have high acuity but a narrow field of view. Between the eyes, there are two unobtrusive central cameras fixed with respect to the head, each with a wider field of view but correspondingly lower acuity. The reason for this mixture of cameras is that typical visual tasks require both high acuity and a wide field of view. High acuity is needed for recognition tasks and for controlling precise visually guided motor movements. A wide field of view is needed for search tasks, for tracking multiple objects, compensating for involuntary ego-motion, etc. A common trade-off found in biological systems is to sample part of the visual field at a high enough resolution to support the first set of tasks, and to sample the rest of the field at an adequate level to support the second set. This is seen in animals with foveate vision, such as humans, where the density of photoreceptors is highest at the center and falls off dramatically towards the periphery. This can be implemented by using specially designed imaging hardware [7], space-variant image sampling [8], or by using multiple cameras with different fields of view, as we have done.

The designs of our robots are constantly evolving. New degrees of freedom are added, old degrees of freedom are reorganized, sensors are replaced or rearranged, new sensory modalities are introduced. The descriptions given here should be treated as a fleeting snapshot of the current state of the robots. Our hardware and software control architectures have been designed to meet the challenge of real-time processing of visual signals (approaching 30 Hz) with minimal latencies. Kismet's vision system is implemented on a network of nine 400 MHz commercial PCs running the QNX real-time operating system. Kismet's motivational system runs on a collection of four Motorola 68332 processors. Machines running Windows NT and Linux are also networked for speech generation and recognition respectively. Even more so than Kismet's physical form, the control network is rapidly evolving as new behaviors and sensory modalities come on line.

IV. Levels of visual behavior

Visual behavior can be conceptualized on four different levels (as shown in Figure 3). These levels correspond to the *social level*, the *behavior level*, the *skills level*, and the *primitives level*. This decomposition is motivated by distinct temporal, perceptual, and interaction constraints that exist at each level. The temporal constraints pertain to how fast the motor acts must be updated and executed. These can range from real-time vision rates (30 Hz) to the relatively slow time scale of social interaction (potentially transitioning over minutes). The perceptual constraints pertain to what level of sensory feedback is required to coordinate behavior at that layer. This perceptual feedback can originate from the low level visual processes such as the current target from the attention system, to relatively high-level multi-modal percepts generated by the behavioral releasers. The interaction constraints pertain to the arbitration of units that compose each layer. This can range from low-level oculomotor primitives (such as saccades and smooth pursuit), to using visual behavior to regulate human-robot turn-taking.

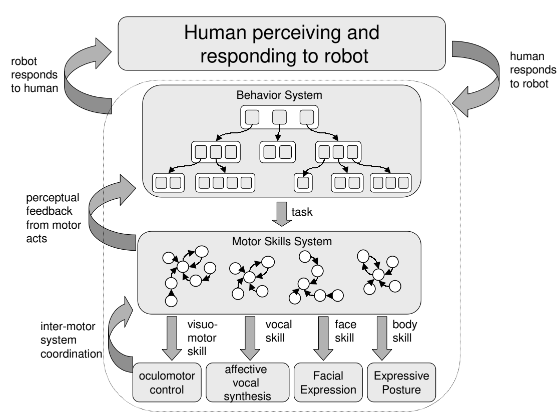

Fig. 3. Levels of behavioral organization. The primitive level is populated with tightly coupled sensorimotor loops. The skill level contains modules that coordinate primitives to achieve tasks. Behavior level modules deal with questions of relevance, persistence and opportunism in the arbitration of tasks. The social level comprises design-time considerations of how the robot's behaviors will be interpreted and responded to in a social environment.

Each level serves a particular purpose for generating the overall observed behavior. As such, each level must address a specific set of issues. The levels of abstraction help simplify the overall control of visual behavior by restricting each level to address those core issues that are best managed at that level. By doing so, the coordination of visual behavior at each level (i.e., arbitration), between the levels (i.e., top-down and bottom-up), and through the world is maintained in a principled way.

- *The Social Level:* The social level explicitly deals with issues pertaining to having a human in the interaction loop. This requires careful consideration of how the human interprets and responds to the robot's behavior in a social context. Using visual behavior (making eye contact and breaking eye contact) to help regulate the transition of speaker turns during vocal turn-taking is an example.
- *The behavior level:* The behavior level deals with issues related to producing relevant, appropriately persistent, and opportunistic behavior. This involves arbitrating between the many possible goal-achieving behaviors that the robot could perform to establish the current task. Actively seeking out a desired stimulus and then visually engaging it is an example.
- *The motor skill level:* The motor skill level is responsible for figuring out how to move the motors to accomplish that task. Fundamentally, this level deals with the issues of blending of and sequencing between coordinated ensembles of motor primitives (each ensemble is a distinct motor skill). The skills level must also deal with coordinating multi-modal motor skills (e.g., those motor skills that combine speech, facial expression, and body posture). Fixed action patterns such as a searching behavior is an example where the robot alternately performs ballistic eye-neck orientation movements with gaze fixation to the most salient target. The ballistic movements are important for scanning the scene, and the fixation periods are important for locking on the desired type of stimulus.
- *The motor primitives level:* The motor primitives level implements the building blocks of motor action. This level must deal with motor resource allocation and tightly coupled sensorimotor loops. For example, gaze stabilization must take sensory stimuli and produce motor commands in a very tight feedback loop. Kismet actually has four distinct motor systems at the primitives level: the *affective vocal system*, the *facial expression system*, the *oculomotor system*, and the *body posturing system*. Because this paper focuses on visual behavior, we only discuss the oculomotor system here.

We describe these levels in detail as they pertain to Kismet's visual behavior. We begin at the lowest level, motor primitives pertaining to vision, and progress to the highest level where we discuss the social constraints of animate vision.

V. Visual motor primitives

Kismet's visual-motor control is modeled after the human ocular-motor system. The human system is so good at providing a stable percept of the world that we have no intuitive appreciation of the physical constraints under which it operates. Humans have foveate vision. The fovea (the center of the retina) has a much higher density of photoreceptors than the periphery. This means that to see an object clearly, humans must move their eyes such that the image of the object falls on the fovea. Human eye movement is not smooth. It is composed of many quick jumps, called saccades, which rapidly re-orient the eye to project a different part of the visual scene onto the fovea. After a saccade, there is typically a period of fixation, during which the eyes are relatively stable. They are by no means stationary, and continue to engage in corrective micro-saccades and other small movements. If the eyes fixate on a moving object, they can follow it with a continuous tracking movement called smooth pursuit. This type of eye movement cannot be evoked voluntarily, but only occurs in the presence of a moving object. Periods of fixation typically end after some hundreds of milliseconds, after which a new saccade will occur [9].

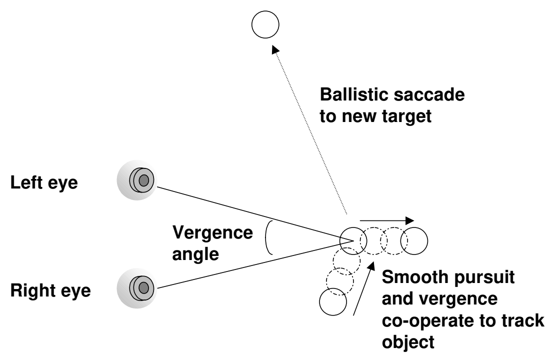

Fig. 4. Humans exhibit four characteristic types of eye motion. Saccadic movements are high-speed ballistic motions that center a target in the field of view. Smooth pursuit movements are used to track a moving object at low velocities. The vestibulo-ocular and opto-kinetic reflexes act to maintain the angle of gaze as the head and body move through the world. Vergence movements serve to maintain an object in the center of the field of view of both eyes as the object moves in depth.

The eyes normally move in lock-step, making equal, conjunctive movements. For a close object, the eyes need to turn towards each other somewhat to correctly image the object on the foveae of the two eyes. These disjunctive movements are called vergence, and rely on depth perception (see Figure 4). Since the eyes are located on the head, they need to compensate for any head movements that occur during fixation. The vestibulo-ocular reflex uses inertial feedback from the vestibular system to keep the orientation of the eyes stable as the eyes move. This is a very fast response, but is prone to the accumulation of error over time. The opto-kinetic response is a slower compensation mechanism that uses a measure of the visual slip of the image across the retina to correct for drift. These two mechanisms work together to give humans stable gaze as the head moves.

Our implementation of an ocular-motor system is an approximation of the human system. The motor primitives are organized around the needs of higher levels, such as maintaining and breaking mutual regard, performing visual search, etc. Since our motor primitives are tightly bound to visual attention, we will first discuss their sensory component.

VI. Pre-attentive visual perception

Human infants and adults naturally find certain perceptual features interesting. Features such as color, motion, and face-like shapes are very likely to attract our attention [10]. We have implemented a variety of perceptual feature detectors that are particularly relevant to interacting with people and objects. These include low-level feature detectors attuned to quickly moving objects, highly saturated color, and colors representative of skin tones. Examples of features we have used are shown in Figure 5. Looming objects are also detected pre-attentively, to facilitate a fast reflexive withdrawal.

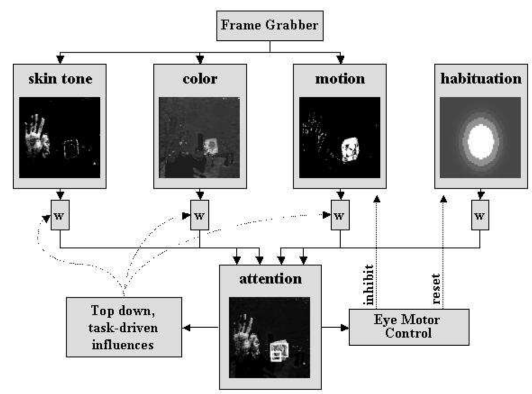

Fig. 5. Overview of the attention system. The robot's attention is determined by a combination of low-level perceptual stimuli. The relative weightings of the stimuli are modulated by high-level behavior and motivational influences. A sufficiently salient stimulus in any modality can pre-empt attention, similar to the human response to sudden motion. All else being equal, larger objects are considered more salient than smaller ones. The design is intended to keep the robot responsive to unexpected events, while avoiding making it a slave to every whim of its environment. With this model, people intuitively provide the right cues to direct the robot's attention (shake object, move closer, wave hand, etc.). Displayed images were captured during a behavioral trial session.

A. Color saliency feature map

One of the most basic and widely recognized visual feature is color. Our models of color saliency are drawn from the complementary work on visual search and attention from Itti, Koch, and Niebur [11]. The incoming video stream contains three 8-bit color channels (r, g, and b) which are transformed into four color-opponency channels (r', g', b', and y'). Each input color channel is first normalized by the luminance l (a weighted average of the three input color channels):

$$r_n = \frac{255}{3} \cdot \frac{r}{l} \qquad g_n = \frac{255}{3} \cdot \frac{g}{l} \qquad b_n = \frac{255}{3} \cdot \frac{b}{l} \tag{1}$$

These normalized color channels are then used to produce four opponent-color channels:

$$r' = r_n - (g_n + b_n)/2 \tag{2}$$

$$g' = g_n - (r_n + b_n)/2 \quad (3)$$
$$b' = b_n - (r_n + g_n)/2 \quad (4)$$
$$y' = \frac{r_n + g_n}{2} - b_n - \|r_n - g_n\| \quad (5)$$

The four opponent-color channels are clamped to 8-bit values by thresholding. While some research seems to indicate that each color channel should be considered individually [10], we choose to maintain all of the color information in a single feature map to simplify the processing requirements (as does Wolfe [12] for more theoretical reasons).

B. Motion feature map

In parallel with the color saliency computations, a second processor receives input images from the frame grabber and computes temporal differences to detect motion. Motion detection is performed on the wide FoV camera, which is often at rest since it does not move with the eyes. The incoming image is converted to grayscale and placed into a ring of frame buffers. A raw motion map is computed by passing the absolute difference between consecutive images through a threshold function $\mathcal{T}$:

$$M_{raw} = \mathcal{T}(\|I_t - I_{t-1}\|) \quad (6)$$

This raw motion map is then smoothed with a uniform 7×8 field. The result is a binary 2-D map where regions corresponding to motion have a high intensity value. The motion saliency feature map is computed at 25-30 Hz by a single 400MHz processor node. Figure 5 gives an example of the motion feature map when the robot looks at a toy block that is being shaken.

C. Skin tone feature map

Colors consistent with skin are also filtered for. This is a computationally inexpensive means to rule out regions that are unlikely to contain faces or hands. Most pixels on faces will pass these tests over a wide range of lighting conditions and skin color. Pixels that pass these tests are weighted according to a function learned from instances of skin tone from images taken by Kismet's cameras (see figure 6). In this implementation, a pixel is *not* skin-toned if:

- $r < 1.1 \cdot g$ (the red component fails to dominate green sufficiently)
- $r < 0.9 \cdot b$ (the red component is excessively dominated by blue)
- $r > 2.0 \cdot max(g, b)$ (the red component completely dominates both blue and green)
- $r < 20$ (the red component is too low to give good estimates of ratios)
- $r > 250$ (the red component is too saturated to give a good estimate of ratios)

VII. Visual attention

We have implemented Wolfe's model of human visual search and attention [12]. Our implementation is similar

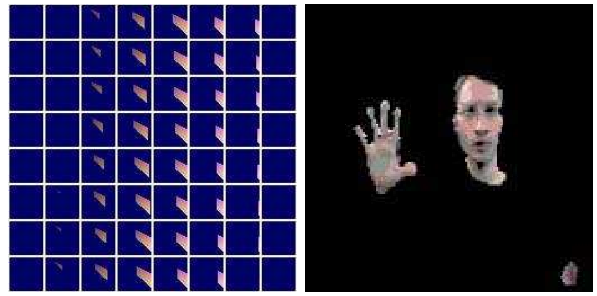

Fig. 6. The skin tone filter responds to 4.7% of possible (R, G, B) values. Each grid in the figure to the left shows the response of the filter to all values of red and green for a fixed value of blue. The image to the right shows the filter in operation. Typical indoor objects that may also be consistent with skin tone include wooden doors, cream walls, etc.

to other models based in part on Wolfe's work [11], but additionally operates in conjunction with motivational and behavioral models, with moving cameras, and addresses the issue of habituation. The attention process acts in two parts. The low-level feature detectors discussed in the previous section are combined through a weighted average to produce a single attention map. This combination allows the robot to select regions that are visually salient and to direct its computational and behavioral resources towards those regions. The attention system also integrates influences from the robot's internal motivational and behavioral systems to bias the selection process. For example, if the robot's current goal is to interact with people, the attention system is biased toward objects that have colors consistent with skin tone. The attention system also has mechanisms for habituating to stimuli, thus providing the robot with a primitive attention span. The state of the attention system is usually reflected in the robot's gaze direction, unless there are behavioral reasons for this not to be the case. The attention system runs all the time, even when it is not controlling gaze, since it determines the perceptual input to which the motivational and behavioral systems respond.

A. Task-based influences on attention

For a goal achieving creature, the behavioral state should also bias what the creature attends to next. For instance, when performing visual search, humans seem to be able to preferentially select the output of one broadly tuned channel per feature (e.g. "red" for color and "shallow" for orientation if searching for red horizontal lines).

In our system these top-down, behavior-driven factors modulate the output of the individual feature maps before they are summed to produce the bottom-up contribution. This process selectively enhances or suppresses the contribution of certain features, but does not alter the underlying raw saliency of a stimulus. To implement this, the bottom-up results of each feature map are passed through a filter (effectively a gain). The value of each gain is determined by the active behavior. For instance, as shown in Figure 7, the skin tone gain is enhanced when the `seek people` behavior is active and is suppressed when the `avoid people` behavior is active. Similarly, the color gain is enhanced

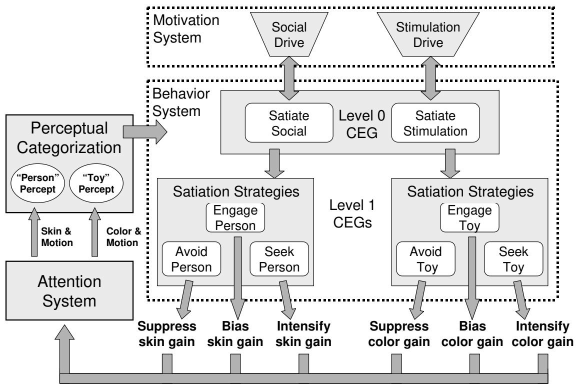

Fig. 7. Schematic of behaviors relevant to attention. The activation of a particular behavior depends on both perceptual factors and motivation factors. The perceptual factors come from post attentive processing of the target stimulus into behaviorally relevant percepts. The drives within the motivation system have an indirect influence on attention by influencing the behavioral context. The behaviors at Level 1 of the behavior system directly manipulate the gains of the attention system to benefit their goals. Through behavior arbitration, only one of these behaviors is active at any time. These behaviors are further elaborated in deeper levels of the behavior system.

when the `seek toys` behavior is active, and suppressed when the `avoid toys` behavior is active. Whenever the `engage people` or `engage toys` behaviors are active, the face and color gains are restored to their default values, respectively.

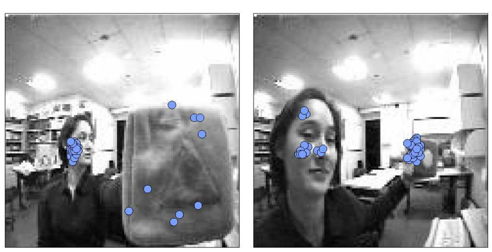

Fig. 8. Effect of gain adjustment on looking preference. Circles correspond to fixation points, sampled at one second intervals. On the left, the gain of the skin tone filter is higher. The robot spends more time looking at the face in the scene (86% face, 14% block). This bias occurs despite the fact that the face is dwarfed by the block in the visual scene. On the right, the gain of the color saliency filter is higher. The robot now spends more time looking at the brightly colored block (28% face, 72% block).

These modulated feature maps are then summed to compute the overall attention activation map, thus biasing attention in a way that facilitates achieving the goal of the active behavior. For example, if the robot is searching for social stimuli, it becomes sensitive to skin tone and less sensitive to color. Behaviorally, the robot may encounter toys in its search, but will continue until a skin toned stimulus is found (often a person's face). Figure 8 shows the results of two such experiments. The left figure shows a looking preference to a person despite a lesser "raw" saliency when the robot is seeking out people. Conversely, when the robot is actively searching for a toy, it demonstrates a looking preference to the colorful block depite the dominant presence of a person's face in the visual field.

B. *Habituation effects*

To build a believable creature, the attention system must also implement habituation effects. Infants respond strongly to novel stimuli, but soon habituate and respond less as familiarity increases. This acts both to keep the infant from being continually fascinated with any single object and to force the caretaker to continually engage the infant with slightly new and interesting interactions. For a robot, a habituation mechanism removes the effects of highly salient background objects that are not currently involved in direct interactions as well as placing requirements on the caretaker to maintain interaction with slightly novel stimulation.

To implement habituation effects, a *habituation filter* is applied to the activation map over the location currently being attended to. The habituation filter effectively decays the activation level of the location currently being attended to, making other locations of lesser activation bias attention more strongly.

C. *Consistency of attention*

In the presence of objects of similar salience, it is useful be able to commit attention to one of the objects for a period of time. This gives time for post-attentive processing to be carried out on the object, and for downstream processes to organize themselves around the object. As soon as a decision is made that the object is not behaviorally relevant (for example, it may lack eyes, which are searched for post-attentively), attention can be withdrawn from it and visual search may continue. Committing to an object is also useful for behaviors that need to be atomically applied to a target (for example, a calling behavior where the robot needs to stay looking at the person it is calling).

To allow such commitment, the attention system is augmented with a tracker. The tracker follows a target in the visual field, using simple correlation between successive frames. Usually changes in the tracker target will be reflected in movements of the robot's eyes, unless this is behaviorally inappropriate. If the tracker loses the target, it has a very good chance of being able to reacquire it from the attention system.

D. *Experiments with directing attention*

The overall attention system runs at 20 Hz on several 400 MHz processors. In section VII-A, we presented Kismet's looking preference results with respect to directing its attention to task-relevant stimuli. In this section, we examine how easy it is for people to direct the Kismet's attention to a specific target stimulus, and to determine when they have been successful in doing so.

Three naive subjects were invited to interact with Kismet. The subjects ranged in age from 25 to 28 years old. All used computers frequently but were not computer scientists by training. All interactions were video-recorded.

The robot's attention gains were set to their default values so that there would be no strong preference for one saliency feature over another. The subjects were asked to direct the robot's attention to each of the target stimuli. There were seven target stimuli used in the study. Three were saturated color stimuli, three were skin-toned stimuli, and the last was a pure motion stimulus. Each target stimulus was used more than once per subject. These are listed below:

- A highly saturated colorful block
- A bright yellow stuffed dinosaur with multi-color spines
- A bright green cylinder
- A bright pink cup (which is actually detected by the skin tone feature map)
- The person's face
- The person's hand
- A black and white plush cow (which is only salient when moving)

The video was later analyzed to determine which cues the subjects used to attract the robot's attention, which cues they used to determine when they had been successful, and the length of time required to do so. They were also interviewed at the end of the session about which cues they used, which cues they read, and about how long they thought it took to direct the robot's attention. The results are summarized in Table I.

stimulus category	stimulus	presentations	average time (s)	commonly used cues	commonly read cues
color & movement	yellow dinosaur	8	8.5	motion across center line shaking motion bringing target close to robot	eye behavior, especially tracking facial expression, especially raised eyebrows body posture, especially forward lean or withdraw
	multi-colored block	8	6.5		
	green cylinder	8	6.0		
motion only	b/w cow	8	5.0		
skin-toned & movement	pink cup	8	6.5		
	hand	8	5.0		
	face	8	3.5		
Total		56	5.8		

TABLE I
Summary from attention manipulation interactions.

To attract the robot's attention, the most frequently used cues include bringing the target close and in front of the robot's face, shaking the object of interest, or moving it slowly across the centerline of the robot's face. Each cue increases the saliency of a stimulus by making it appear larger in the visual field, or by supplementing the color or skin-tone cue with motion. Note that there was an inherent competition between the saliency of the target and the subject's own face as both could be visible from the wide FoV camera. If the subject did not try to direct the robot's attention to the target, the robot tended to look at the subject's face.

The subjects also effortlessly determined when they had successfully re-directed the robot's gaze. Interestingly, it is not sufficient for the robot to orient to the target. People look for a *change* in visual behavior, from ballistic orientation movements to smooth pursuit movements, before concluding that they had successfully re-directed the robot's attention. All subjects reported that eye movement was the most relevant cue to determine if they had successfully directed the robot's attention. They all reported that it was easy to direct the robot's attention to the desired target. They estimated the mean time to direct the robot's attention at 5 to 10 seconds. This turns out to be the case; the mean time over all trials and all targets is 5.8 seconds.

VIII. Post-attentive processing

Once the attention system has selected regions of the visual field that are potentially behaviorally relevant, more intensive computation can be applied to these regions than could be applied across the whole field. Searching for eyes is one such task. Locating eyes is important to us for engaging in eye contact, and as a reference point for interpreting facial movements and expressions. We currently search for eyes after the robot directs its gaze to a locus of attention, so that a relatively high resolution image of the area being searched is available from the foveal cameras (see Figure 9). Once the target of interest has been selected, we also estimate its proximity to the robot using a stereo match between the two central wide cameras. Proximity is an important for interaction as things closer to the robot should be of greater interest. It's also useful for interaction at a distance, such as a person standing too far for face to face interaction but is close enough to be beckoned closer. Clearly the relevant behavior (beckoning or playing) is dependent on the proximity of the human to the robot.

Fig. 9. Eyes are searched for within a restricted part of the robot's field of view. The eye detector actually looks for the region between the eyes. It has adequate performance over a limited range of distances and face orientations.

Eye-detection in a real-time, robotic domain is computationally expensive and prone to error due to the large variance in head posture, lighting conditions and feature scales. We developed an approach based on successive feature extraction, combined with some inherent domain constraints, to achieve a robust and fast eye-detection system for Kismet. First, a set of feature filters are applied successively to the image in increasing feature granularity. This serves to reduce the computational overhead while maintaining a robust system. The successive filter stages are:

- Detect skin colored patches in the image (abort if this does not pass above threshold).
- Scan the image for ovals and characterize its skin tone for a potential face.
- Extract a sub-image of the oval and run a ratio template [13], [14] over it for candidate eye locations.
- For each candidate eye location, run a pixel based multilayer perceptron on the region. The perceptron is previously trained to recognize shading patterns characteristic of the eyes and bridge of the nose.

By doing so, the set of possible eye-locations in the image is reduced from the previous level based on a feature filter. This allows the eye detector to run in real time on a 400Mhz PC. The methodology assumes that the lighting conditions allow the eyes to be distinguished as dark regions surrounded by highlights of the temples and the bridge of the nose, that human eyes are largely surrounded by regions of skin color, that the head is only moderately rotated, that the eyes are reasonably horizontal, and that people are within interaction distance from the robot (3 to 10 feet).

IX. Eye movements

Kismet's eyes periodically saccade to new targets chosen by an attention system, tracking them smoothly if they move and the robot wishes to engage them. Vergence eye movements are more challenging to implement in a social setting, since errors in disjunctive eye movements can give the eyes a disturbing appearance of moving independently. Errors in conjunctive movements have a much smaller impact on an observer, since the eyes clearly move in lockstep. A crude approximation of the opto-kinetic reflex is rolled into our implementation of smooth pursuit. An analogue of the vestibular-ocular reflex has been developed using a 3-axis inertial sensor, but has yet to be implemented on Kismet (it currently runs on other humanoid robots in our lab). Kismet uses an efferent copy mechanism to compensate the eyes for movements of the head. An over of the oculomotor control system is shown in figure 10.

The attention system operates on the view from the central camera (see Figure 2). A transformation is needed to convert pixel coordinates in images from this camera into position setpoints for the eye motors. This transformation in general requires the distance to the target to be known, since objects in many locations will project to the same point in a single image (see Figure 11). Distance estimates are often noisy, which is problematic if the goal is to center the target exactly in the eyes. In practice, it is usually enough to get the target within the field of view of the foveal cameras in the eyes. Clearly the narrower the field of view of these cameras is, the more accurately the distance to the object needs to be known. Other crucial factors are the distance between the wide and foveal cameras, and the closest distance at which the robot will need to interact with objects. These constraints determined the physical distribution of Kismet's cameras and choice of lenses. The central location of the wide camera places it as close as possible to the foveal cameras. It also has the advantage that moving the head to center a target as seen in the central camera will in fact truly orient the head towards that target – for cameras in other locations, accuracy of orientation would be limited by the accuracy of the measurement of distance.

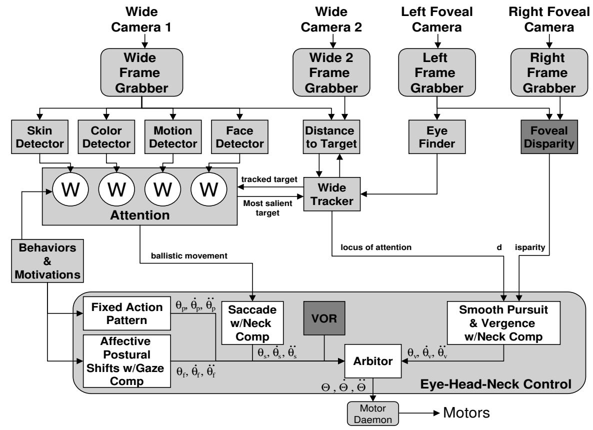

Fig. 10. Organization of Kismet's eye/neck motor control. Many cross level influences have been omitted. The modules in gray are not active in the results presented in this paper.

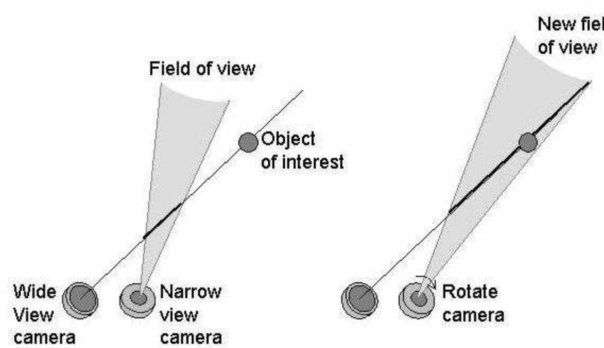

Fig. 11. Without distance information, knowing the position of a target in the wide camera only identifies a ray along which the object must lie, and does not uniquely identify its location. If the cameras are close to each other relative to the closest distance the object is expected to be at, the foveal cameras can be rotated to bring the object within their narrow field of view without needing an accurate estimate of its distance. If the cameras are far apart, or the field of view is very narrow, the minimum distance the object can be at becomes large.

Higher-level influences modulate eye and neck movements in a number of ways. As already discussed, modifications to weights in the attention system translate to changes of the locus of attention about which eye movements are organized. The regime used to control the eyes and neck is available as a set of primitives to higher-level modules. Regimes include low-commitment search, high-commitment engagement, avoidance, sustained gaze, and deliberate gaze breaking. The primitive percepts generated by this level include a characterization of the most salient regions of the image in terms of the feature maps, an extended characterization of the tracked region in terms of the results of post-attentive processing (eye detection, distance estimation), and signals related to undesired conditions, such as a looming object, or an object moving at speeds the tracker finds difficult to keep up with.

X. Visual motor skills

Given the current task (as dictated by the behavior system), the motor skills level is responsible for figuring out how to move the actuators to carry out the stated goal. Often this requires coordination between multiple motor modalities (speech, body posture, facial display, and gaze control). Requests for these modalities can originate from the top-down (e.g. from the emotion system or behavior system), as well as from the bottom-up (the vocal system requesting lip and jaw movements for lip synching). Hence, the motor skills level must address the issue of servicing the motor requests of different systems across the different motor resources.

Furthermore, it often requires a sequence of coordinated motor movements to satisfy a goal. Each motor movement is a primitive (or a combination of primitives) from one of the base motor systems (the vocal system, the oculomotor system, etc.). Each of these coordinated series of motor primitives is called a *skill*, and each skill is implemented as a finite state machine (FSM). Each motor skill encodes knowledge of how to move from one motor state to the next, where each sequence is designed to bring the robot closer to the current goal. The motor skills level must arbitrate among the many different FSMs, selecting the one to become active based on the active goal. This decision process is straight forward since there is an FSM tailored for each task of the behavior system.

Many skills can be thought of as a *fixed action pattern* (FAP) as conceptualized by early ethologists. Each FAP consists of two components, the *action* component and the *taxis* (or orienting) component. For Kismet, FAPs often correspond to communicative gestures where the action component corresponds to the facial gesture, and the taxis component (to whom the gesture is directed) is controlled by gaze. People seem to intuitively understand that when Kismet makes eye contact with them, they are the locus of Kismet's attention and the robot's behavior is organized about them. This places the person in a state of action readiness where they are poised to respond to Kismet's gestures.

A simple example of a motor skill is Kismet's "calling" FAP. When the current task is to bring a person into a good interaction distance, the motor skill system activates the `calling` FSM. The taxis component of the FAP issues a `hold gaze` request to the oculomotor system. This serves to maintain the robot's gaze on the person to be hailed. In the first state of the gestural component, Kismet leans its body toward the person (a request to the body posture motor system). This strengthens the person's perception that the robot has taken a particular interest in them. The ears also begin to waggle exuberantly (creating a significant amount of motion and noise) which further attracts the person's attention to the robot (a request to the face motor system). In addition, Kismet vocalizes excitedly which is perceived as an initiation of engagement. At the completion of this gesture, the FSM transitions to the second state. In this state, the robot "sits back" and waits for a bit with an expecting expression (ears slightly perked, eyes slightly widened, and brows raised). If the person has not already approached the robot, it is likely to occur during this "anticipation" phase. If the person does not approach within the allotted time period, the FSM transitions to the third state in which the face relaxes, the robot maintains a neutral posture, and gaze fixation is released. At this point, the robot is likely to shift gaze. As long as this FSM is active (determined by the behavior system), the hailing cycle repeats. It can be interrupted at any state transition by the activation of another FSM (such as the "greeting" FSM when the person has approached).

We now move up another layer of abstraction, to the behavior level in the hierarchy that was shown in Figure 3.

XI. Visual behavior

The behavior level is responsible for establishing the current task for the robot through arbitrating among Kismet's goal-achieving behaviors. By doing so, the observed behavior should be relevant, appropriately persistent, and opportunistic. Both the current environmental conditions (as characterized by high-level perceptual releasers, as well as motivational factors (emotion processes and homeostatic regulation) contribute to this decision process.

Interaction of the behavior level with the social level occurs through the world as determined by the nature of the interaction between Kismet and the human. As the human responds to Kismet, the robot's perceptual conditions change. This can activate a different behavior, whose goal is physically carried out the underlying motor systems. The human observes the robot's ensuing response and shapes their reply accordingly.

Interaction of the behavior level with the motor skills level also occurs through the world. For instance, if Kismet is looking for a bright toy, then the `seek toy` behavior is active. This task is passed to the underlying motor skills which carry out the search. The act of scanning the environment brings new perceptions to Kismet's field of view. If a toy is found, then the `seek toy` behavior is successful and released. At this point, the perceptual conditions for engaging the toy are relevant and the `engage toy` behaviors become active. A new set motor skills become active to track and smoothly pursue the toy.

A. Social level

Eye movements have communicative value. As discussed previously, they indicate the robot's locus of attention. The robot's degree of engagement can also be conveyed, to communicate how strongly the robot's behavior is organized around what it is currently looking at. If the robot's eyes flick about from place to place without resting, that indicates a low level of engagement, appropriate to a visual search behavior. Prolonged fixation with smooth pursuit and orientation of the head towards the target conveys a much greater level of engagement, suggesting that the robot's behavior is very strongly organized about the locus of attention.

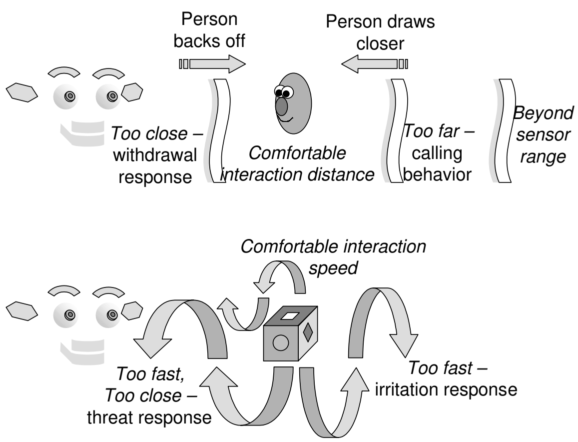

Fig. 12. Regulating interaction. People too distant to be seen clearly are called closer; if they come too close, the robot signals discomfort and withdraws. The withdrawal moves the robot back somewhat physically, but is more effective in signaling to the human to back off. Toys or people that move too rapidly cause irritation.

Eye movements are the most obvious and direct motor actions that support visual perception. But they are by no means the only ones. Postural shifts and fixed action patterns involving the entire robot also have an important role. Kismet has a number of coordinated motor actions designed to deal with various limitations of Kismet's visual perception (see Figure 12). For example, if a person is visible, but is too distant for their face to be imaged at adequate resolution, Kismet engages in a calling behavior to summon the person closer. People who come too close to the robot also cause difficulties for the cameras with narrow fields of view, since only a small part of a face may be visible. In this circumstance, a withdrawal response is invoked, where Kismet draws back physically from the person. This behavior, by itself, aids the cameras somewhat by increasing the distance between Kismet and the human. But the behavior can have a secondary and greater effect through social amplification – for a human close to Kismet, a withdrawal response is a strong social cue to back away, since it is analogous to the human response to invasions of "personal space."

Similar kinds of behavior can be used to support the visual perception of objects. If an object is too close, Kismet can lean away from it; if it is too far away, Kismet can crane its neck towards it. Again, in a social context, such actions have power beyond their immediate physical consequences. A human, reading intent into the robot's actions, may amplify those actions. For example, neck-craning towards a toy may be interpreted as interest in that toy, resulting in the human bringing the toy closer to the robot. Another limitation of the visual system is how quickly it can track moving objects. If objects or people move at excessive speeds, Kismet has difficulty tracking them continuously. To bias people away from excessively boisterous behavior in their own movements or in the movement of objects they manipulate, Kismet shows irritation when its tracker is at the limits of its ability. These limits are either physical (the maximum rate at which the eyes and neck move), or computational (the maximum displacement per frame from the cameras over which a target is searched for).

Such regulatory mechanisms play roles in more complex social interactions, such as conversational turn- taking. Here control of gaze direction is important for regulating conversation rate [15]. In general, people are likely to glance aside when they begin their turn, and make eye contact when they are prepared to relinquish their turn and await a response. Blinks occur most frequently at the end of an utterance. These and other cues allow Kismet to influence the flow of conversation to the advantage of its auditory processing. The visual-motor system can also be driven by the requirements of a nominally unrelated sensory modality, just as behaviors that seem completely orthogonal to vision (such as ear-wiggling during the call behavior to attract a person's attention) are nevertheless recruited for the purposes of regulation. These mechanisms also help protect the robot. Objects that suddenly appear close to the robot trigger a looming reflex, causing the robot to quickly withdraw and appear startled. If the event is repeated, the response quickly habituates and the robot simply appears annoyed, since its best strategy for ending these repetitions is to clearly signal that they are undesirable. Similarly, rapidly moving objects close to the robot are threatening and trigger an escape response. These mechanisms are all designed to elicit natural and intuitive responses from humans, without any special training. But even without these carefully crafted mechanisms, it is often clear to a human when Kismet's perception is failing, and what corrective action would help, because the robot's perception is reflected in behavior in a familiar way. Inferences made based on our human preconceptions are actually likely to work.

B. Evidence of Social Amplification

To evaluate the social implications of Kismet's behavior, we invited a few people to interact with the robot in a free-form exchange. There were four subjects in the study, two males (one adult and one child) and two females (both adults). They ranged in age from 12 to 28 years. None of the subjects were affiliated with MIT. All had substantial experience with computers. None of the subjects had any prior experience with Kismet. The child had prior experience with a variety of interactive toys. Each subject interacted with the robot for 20 to 30 minutes. All exchanges were video recorded for further analysis.

We analyzed the video for evidence of social amplification. Namely, did people read Kismet's cues and did they respond to them in a manner that benefited the robot's perceptual processing or its behavior? we found several classes of interactions where the robot displayed social cues and successfully regulated the exchange.

B.1 Establishing a Personal Space

The strongest evidence of social amplification was apparent in cases where people came within very close proximity of Kismet. In numerous instances the subjects would bring their face very close to the robot's face. The robot would withdraw, shrinking backwards, perhaps with an annoyed expression on its face. In some cases the robot would also issue a vocalization with an expression of disgust. In one instance, the subject accidentally came too close and the robot withdrew without exhibiting any signs of annoyance. The subject immediately queried, "Am I too close to you? I can back up," and moved back to put a bit more space between himself and the robot. In another instance, a different subject intentionally put his face very close to the robot's face to explore the response. The robot withdrew while displaying full annoyance in both face and voice. The subject immediately pushed backwards, rolling the chair across the floor to put about an additional three feet between himself and the robot, and promptly apologized to the robot.

Overall, across different subjects, the robot successfully established a personal space. This benefits the robot's visual processing by keeping people at a distance where the visual system can detect eyes more robustly. This behavioral response was added to the robot's repertoire because previous interactions with naive subjects illustrated the robot was not granted any personal space. This can be attributed to "baby movements" where people tend to get extremely close to infants, for instance.

B.2 Luring People to a Good Interaction Distance

People seem responsive to Kismet's calling behavior. When a person is close enough for the robot to perceive his/her presense, but too far away for face-to-face exchange, the robot issues this social display to bring the person closer. The most distinguishing features of the display are craning the neck forward in the direction of the person, wiggling the ears with large amplitude, and vocalizing with an excited affect. The function of the display is to lure people into an interaction distance that benefits the vision system. This behavior is not often witnessed as most subjects simply pull up a chair in front of the robot and remain seated at a typical face-to-face interaction distance.

The youngest subject took the liberty of exploring different interaction ranges, however. Over the course of about fifteen minutes he would alternately approach the robot to a normal face-to-face distance, move very close to the robot (invading its personal space), and backing away from the robot. Upon the first appearance of the calling response, the experimenter queried the subject about the robot's behavior. The subject interpreted the display as the robot wanting to play, and he approached the robot. At the end of the subject's investigation, the experimenter queried him about the further interaction distances. The subject responded that when he was further from Kismet, the robot would lean forward. He also noted that the robot had a harder time looking at his face when he was farther back. In general, he interpreted the leaning behavior as the robot's attempt to initiate an exchange with him. We have noticed from earlier interactions (with other people unfamiliar with the robot) that a few people have not immediately understood this display as a "calling" behavior. The display is flamboyant enough, however, to arouse their interest to approach the robot.

XII. Limitations and extensions

There are a number of ways the current implementation could be improved and expanded upon. Some of these recommendations involve supplementing the existing framework, others involve integrating this system into a larger framework.

Kismet's visual perceptual world only consists of what is in view of the cameras. Ultimately, the robot should be able to construct an ego-centered saliency map of interaction space. In this representation, the robot could keep track of where interesting things are located, even if they are not currently in view. Human infants engage in social referencing with their caregiver at a very young age. If some event occurs that the infant is unsure about, the infant will look to the caregiver's face for an affective assessment. The infant will use this assessment to organize its behavior. For instance, if the caregiver looks frightened, the infant may become distressed and not probe further. If the caregiver looks pleased and encouraging, the infant is likely to continue exploring. With respect to Kismet, it will encounter many situations that it was not explicitly programmed to evaluate. However, if the robot can engage in social referencing, it can look to the human for the affective assessment and use it to bias learning and to organize subsequent behavior. Chances are, the event in question and the human's face will not be in view at the same time. Hence, a representation of where interesting this are in ego-centered interaction space is an important resource.

The attention system could be extended by adding new feature maps. A depth map from stereo would be very useful – currently distance is only computed post-attentively. Another interesting feature map to incorporate into the system would be edge orientation. Wolfe and Triesman among others argue in favor of edge orientation as a bottom-up feature map in humans. Currently, Kismet has no shape metrics to help it distinguish objects from each other (such as its toy block from its toy dinosaur). Adding features to support this is an important extension to the existing implementation.

There are no auditory bottom-up contributions. A sound localization feature map would be a nice multi-modal extension. Currently, Kismet assumes that the most salient person is the one who is talking to it. Often there are multiple people talking around and to the robot. It is important that the robot knows who is addressing it and when. Sound localization would be of great benefit here.

Acknowledgments

This work was funded by a DARPA under contract DABT 63-99-1-0012 and by NTT.

XIII. Conclusions

Motor control for a social robot poses challenges beyond issues of stability and accuracy. Motor actions will be perceived by human observers as semantically rich, regardless of whether the imputed meaning is intended or not. This can be a powerful resource for facilitating natural interactions between robot and human, and places constraints on the robot's physical appearance and movement. It allows the robot to be readable – to make its behavioral intent and motivational state transparent at an intuitive level to those it interacts with. It allows the robot to regulate its interactions to suit its perceptual and motor capabilities, again in an intuitive way with which humans naturally cooperate. These social constraints give the robot leverage over the world that extends far beyond its physical competence, through social amplification of its perceived intent. If properly designed, the robot's visual behaviors can be matched to human expectations and allow both robot and human to participate in natural and intuitive social interactions.

References

[1] D.H. Ballard, "Animate vision," *AI*, vol. 48, pp. 57–86, 1991.
[2] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, "Overt visual attention for a humanoid robot," in *Proceedings of the International Conference on Intelligence in Robotics and Autonomous Systems (IROS01)*, Hawaii, 2001.
[3] K. Kawamura, A. Alford, K. Hambuchen, and M. Wilkes, "Towards a unified framework for human-humanoid interaction," in *Proceedings of the First IEEE-RAS International Conference on Humanoid Robots*, Cambridge, MA, 2000.
[4] G. Metta, *BabyRoBot: A study in sensori-motor development*, Ph.D. thesis, University of Genoa, LIRA-Lab, DIST, 1999.
[5] H. Kozima, "Attention-sharing and behavior-sharing in human-robot communication," in *IEEE International Workshop on Robot and Human Communication (ROMAN-98)*, Takamatsu, Japan, 1998, pp. 9–14.
[6] D.H. Ballard, "Behavioral constraints on animate vision," *Image and Vision Computing*, vol. 7:1, pp. 3–9, 1989.
[7] Jan van der Spiegel, G. Kreider, C. Claeys, I. Debusschere, G. Sandini, P. Dario, F. Fantini, P. Belluti, and G. Soncini, "A foveated retina-like sensor using ccd technology," in *Analog VLSI Implementation of Neural Systems*, C. Mead and M. Ismail, Eds., pp. pp. 189–212. Kluwer Academic Publishers, 1989.
[8] A. Bernardino and J. Santos-Victor, "Binocular visual rracking: Integration of perception and control," *IEEE Transactions on Robotics and Automation*, vol. 15(6), pp. 1937–1958, 1999.
[9] Michael E. Goldberg, "The control of gaze," in *Principles of Neural Science*, Eric R. Kandel, James H. Schwartz, and Thomas M. Jessell, Eds. McGraw-Hill, 4rd edition, 2000.
[10] H. C. Nothdurft, "The role of features in preattentive vision: Comparison of orientation, motion and color cues," *Vision Research*, vol. 33, pp. 1937–1958, 1993.
[11] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 20, no. 11, pp. 1254–1259, 1998.
[12] Jeremy M. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.
[13] B. Scassellati, "Finding eyes and faces with a foveated vision system," in *Proceedings of the Fifthteenth National Conference on Artificial Intelligence (AAAI98)*, Madison, WI, 1998, pp. 969–976.
[14] Pawan Sinha, "Object recognition via image invariants: A case study," *Investigative Ophthalmology and Visual Science*, vol. 35, pp. 1735–1740, May 1994.
[15] J. Cassell, "Embodied conversation: Integrating face and gesture into automatic spoken dialogue systems," .

Cynthia Breazeal is a postdoctoral associate at MIT. She received her Sc.D. from MIT in the Department of Electrical Engineering and Computer Science in May 2000. Her interests focus on human-like robots that can interact in natural, social ways with humans. She received her BS in Electrical and Computer Engineering at University of California, Santa Barbara, and her MS in Electrical Engineering and Computer Science at MIT.

Aaron Edsinger is currently a graduate student with Professor Rodney Brooks at the MIT Artificial Intelligence Lab. He is interested in combining sophisticated sensorimotor control with aesthetic considerations to create naturalistic movement in robots. He received a BS in Computer Systems at Stanford.

Paul Fitzpatrick recieved a B.Eng. and M.Eng. in Computer Engineering at the University of Limerick, Ireland. He is currently a graduate student with Prof. Rodney Brooks a the MIT Artificial Intelligence Laboratory. His current work focuses on robot-human communication through social protocols realized in the visual and auditory domains.

Brian Scassellati is completing his PhD with Rodney Brooks at the MIT Artificial Intelligence Lab. His work is strongly grounded in theories of how the human mind develops, and he is interested in robotics as a tool for evaluating models from biological sciences. He received a BS in Computer Science, a BS in Brain and Cognitive Science, and his ME in Electrical Engineering and Computer Science, all from MIT.